Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers

Douglas S. Brungart[a)]
Air Force Research Laboratory, Human Effectiveness Directorate, 2610 Seventh Street, Wright-Patterson AFB, Ohio 45433

Peter S. Chang
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210

Brian D. Simpson
Air Force Research Laboratory, Human Effectiveness Directorate, 2610 Seventh Street, Wright-Patterson AFB, Ohio 45433

DeLiang Wang
Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

(Received 11 December 2007; revised 6 March 2009; accepted 23 March 2009)

When a target voice is masked by an increasingly similar masker voice, increases in energetic masking are likely to occur due to increased spectro-temporal overlap in the competing speech waveforms. However, the impact of this increase may be obscured by informational masking effects related to the increased confusability of the target and masking utterances. In this study, the effects of target-masker similarity and the number of competing talkers on the energetic component of speech-on-speech masking were measured with an ideal time-frequency segregation (ITFS) technique that retained all the target-dominated time-frequency regions of a multitalker mixture but eliminated all the time-frequency regions dominated by the maskers. The results show that target-masker similarity has a small but systematic impact on energetic masking, with roughly a 1 dB release from masking for same-sex maskers versus same-talker maskers and roughly an additional 1 dB release from masking for different-sex masking voices. The results of a second experiment measuring ITFS performance with up to 18 interfering talkers indicate that energetic masking increased systematically with the number of competing talkers. These results suggest that energetic masking differences related to target-masker similarity have a much smaller impact on multitalker listening performance than energetic masking effects related to the number of competing talkers in the stimulus and non-energetic masking effects related to the confusability of the target and masking voices. [DOI: 10.1121/1.3117686]

PACS number(s): 43.71.Gv, 43.66.Dc [RLF] Pages: 4006–4022

I. INTRODUCTION

Speech perception in multitalker listening environments is limited by two very different types of masking. The first is energetic masking, which occurs when the acoustic elements of the target and masking signals overlap in time and frequency in such a way that portions of the target signal are rendered undetectable in the combined audio mixture. The second type of masking, often referred to as informational masking, occurs when the acoustic characteristics of the target and masking voices are perceptually similar and the listener is unable to successfully extract or segregate acoustically detectable target information from the multitalker mixture (Brungart, 2001; Carhart and Tillman, 1969; Freyman *et al.*, 1999; Kidd *et al.*, 1998; Pollack, 1975).

Because both types of masking occur in all multitalker listening tasks, it is often very difficult to tease apart the relative impacts that energetic and informational masking have in even the simplest multitalker listening environments. For example, consider the case of a target talker who is masked either by a similar-sounding talker of the same sex or a very different sounding talker of the opposite sex. Intuitively, it seems quite obvious that performance would be better with a different-sex masker than with a same-sex masker, and many experiments have shown this to be the case (Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Assman and Summerfield, 1994; Darwin *et al.*, 2003; Brungart *et al.*, 2001; Festen and Plomp, 1990). However, the underlying reason for this improvement in performance is not at all easy to identify. In part, it could be the result of reduced spectro-temporal overlap between the target and masking signals. Female voices typically have F0 values about one octave higher than those of males, and they have formant values roughly 16% higher than those of typical

a)Author to whom correspondence should be addressed. Electronic mail: douglas.brungart@wpafb.af.mil

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
06 MAR 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory,Human Effectiveness Directorate,2610 Seventh Street,Wright-Patterson AFB,OH,45433**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**17**	

male talkers (Peterson and Barney, 1952). This should lead to some reduction in spectro-temporal overlap between male and female voices, and thus an improvement in intelligibility due to a reduction in energetic masking.

However, it is quite likely that some improvement in intelligibility will also occur simply because the different-sex talkers are much more distinct and less confusable than same-sex talkers. The difficulty lies in determining the relative contributions that reduced spectral overlap and reduced confusability make to the improvement in performance that occurs with a different-sex masker versus a same-sex masker. The most direct results addressing this issue come from a paper by Festen and Plomp (1990) that compared the speech reception thresholds (SRTs) for speech masked by a same or different-sex talker under a variety of conditions. In the baseline condition, where normal speech was used for both the target and masking voices, the SRT was 7–11 dB lower for a different-sex masking voice than for a same-sex masking voice. However, when the same- or different-sex masking voice was replaced by a noise signal that was shaped to have the same long-term spectrum as the masking talker, the SRT advantage in the different-sex condition dropped to less than 2 dB. This suggests that differences in long-term spectrum can account for very little of the release from masking that occurs with a different-sex masking voice. Of course, such an analysis ignores the fact that the spectra of the target and masking voices vary constantly over the course of the utterance. Festen and Plomp (1990) attempted to capture this effect by dividing the signal into two bands (above and below 1 kHz) and separately modulating the amplitudes of the bands to match the envelopes of the corresponding bands of the original masking speech utterances. Again, the results of this condition showed very little release from masking ($<$2 dB) in the different-sex conditions.

The Festen and Plomp (1990) results suggest that energetic factors related to reduced spectro-temporal overlap can account for very little of the improvement in SRT found for different-sex maskers. However, their technique cannot fully account for the release from energetic masking that is likely to occur with a different-sex masker. At low frequencies, the larger differences in F0 that occur for different-sex talkers might lead to a reduction in the instantaneous overlap in the resolved harmonics of the speech signal that would not be captured in the long-term average spectrum of the masking speech (Qin and Oxenham, 2003). At higher frequencies, the two-band modulation is simply insufficient to capture the true spectral fluctuations of the same and different-sex voices. Increasing the number of modulation bands is problematic as well: as the number of bands increases, the masking speech becomes increasingly speech-like, and there is an ever increasing possibility that performance will be limited by the confusability of the target and masking voices and not by the loss of information incurred due to spectro-temporal overlap in the target and masking waveforms.

Another possible approach for isolating the energetic component of speech on speech masking involves the spatial separation of the target and masking signals. Previous research has shown that spatial separation in the perceived locations of the target and masking signals reduces the potential confusability of the two signals and thus results in a substantial reduction in the informational masking caused by the interfering speech stimulus (Freyman *et al.*, 1999). Earlier studies have also shown that the performance differences between same-sex and different-sex maskers are much smaller when the masking voices are spatially separated from the target voice than when they originate from the same location as the target talker (Brungart and Simpson, 2002). Indeed, in the largest spatial separation conditions tested, where the target was located just outside the listener's right ear and the masker was located 1 m away or vice versa, Brungart and Simpson (2002) found almost no difference in performance with same-sex and difference-sex interfering talkers. If one hypothesizes that the energetic component of speech-on-speech masking from a same- or different-sex talker can be approximated by the amount of residual masking that occurs when a speech masker is perceived to be spatially separated from a target speech signal (Freyman *et al.*, 2007), then this result seems to confirm the results of the Festen and Plomp (1990) study showing that there is only a modest difference in energetic masking between same-sex and different-sex interfering voices.

However, the use of spatial separation as a means to isolate energetic masking effects in multitalker speech stimuli is potentially problematic in two important ways. The first is that it is based on the assumption that informational masking is largely eliminated when two speech signals appear to originate from different locations in space. This assumption contrasts with the results of Arbogast *et al.* (2002), which showed that an interfering speech signal can still produce a substantial amount of informational masking even when it is spatially separated 90° apart from the target speech. That experiment used a sinewave vocoding procedure to produce target and masking speech signals that had little or no spectral overlap but were highly intelligible when they were presented individually. The results of the experiment showed that the spatially-separated speech masker produced roughly 7 dB more masking than a spatially-separated random-phase noise masker with the same spectral content. This 7 dB difference suggests that some residual informational masking may remain even when the perceptual locations of the target and masking talkers are clearly separated.

The second potential problem with the use of spatial separation as a means to isolate energetic masking effects in multitalker listening is that the spatial separation itself will lead to a substantial reduction in the energetic masking component of the stimulus. Spatial separation typically results in an increase in signal-to-noise ratio (SNR) at one of the listeners ears (the so-called "better ear advantage") and an additional release from masking due to low-frequency interaural phase differences in the competing stimuli (the "binaural interaction" effect). Thus, to the extent that spatial separation eliminates the informational component of speech-on-speech masking, it will only provide an indication of the effects of energetic masking for a spatially-separated target-masker pair. In cases where there is a desire to explicitly determine the effects of energetic masking for co-located talkers, or in cases such as those involving monaural listeners where spa-

tial separation of the talkers is not an option, alternative methods of isolating the energetic components of speech-on-speech masking are required.

One possible alternative method for evaluating the impact of spectro-temporal overlap on speech-on-speech masking for same-sex and different-sex masking speech is ideal time-frequency segregation (ITFS), a recently proposed technique that attempts to simulate the effects of energetic masking by removing those spectro-temporal regions of the acoustic mixture where the target signal is dominated by the masker waveform (Brungart *et al.*, 2006). ITFS is a signal processing technique that removes time-frequency (T-F) regions of a mixture where target energy would be rendered undetectable by a more intense masker; at the same time, it retains all the T-F regions of the mixture where the target would remain detectable despite the presence of the masker. The term ITFS comes from the fact that the processed signal represents an "ideal" segregation of the acoustic elements that potentially contain useful information about the target signal from a background that contains only information about the masker. Because it eliminates the T-F regions of the stimulus where the local SNR is negative, the ITFS technique presumably removes the same target information from the stimulus that would ordinarily be irretrievably lost due to spectro-temporal overlap with the masker. The advantage of the technique is that the removal of the masked portions of the stimulus eliminates any acoustic portions of the masker that might be confused with the target and thus cause informational masking. While there are clearly some potential areas (forward/backward masking, T-F resolution, etc.) where the ITFS technique may fail to accurately capture all aspects of energetic masking in speech perception (Brungart *et al.*, 2006), the general framework of the ITFS technique provides a means to examine the energetic masking effects that influence the detection of a target signal in an acoustic mixture without potentially confounding effects that might occur due to informational masking.

Brungart *et al.* (2006) showed that the application of the ITFS technique to an acoustic mixture had very different results on the intelligibility of the target signal for noise maskers and speech maskers. When the masking signal was noise, the application of the ITFS technique resulted in a modest (2–5 dB) improvement in the SRT of the target speech. However, when the masking signal was a 1-, 2-, or 3-talker speech signal, the application of the ITFS technique resulted in a much larger release from masking (22–25 dB). The size of the release from masking obtained for the speech masking stimuli in the 2006 study is an indicator of how effectively the ITFS technique is able to eliminate the speech-on-speech masking component that is related to the potential confusability of the target and masking voices. By comparing the residual masking remaining after the application of ITFS, it should be possible to assess the relative differences in masking that occur due to spectro-temporal overlap in two competing speech signals. Thus, one should be able to use ITFS to determine the extent to which a same-talker masker produces more energetic masking than a same-sex or different-sex masker. However, the 2006 study only used same-talker maskers, so no such comparison is possible.

In this paper, we apply the ITFS technique developed in the 2006 study to examine the impact that differences in target and masker voice characteristics have on the energetic component of masking in multitalker speech perception. In a second experiment, we extend the technique to examine how energetic masking effects vary with the number of competing talkers in the stimulus. Section II describes the ITFS technique in more detail.

II. ITFS

ITFS is closely related to the notion of ideal binary mask originated in computational auditory scene analysis (Wang and Brown, 2006, pp. 22–23). Assuming a two-dimensional T-F representation where elements are called T-F units, an ideal binary mask is defined as a binary matrix where 1 indicates that the target energy in the corresponding T-F unit exceeds the interference energy by a predefined local SNR criterion (LC) and 0 indicates otherwise. The mask is called ideal because its construction requires *a priori* knowledge of the spectral content of the target and masking signals, and the selection of a LC value of 0 dB is known to be optimal in terms of the theoretical SNR gain of the processed output mixture (see Wang, 2005; Li and Wang, 2009).

Figure 1 illustrates the ideal binary masks for two-talker mixtures where the target is a male utterance and the masker is a sentence uttered by the same talker, a different male talker, or a female talker. In the figure, mixtures SNR and LC are both set to 0 dB. The top row of Fig. 1 shows the cochleagram of the target. Similar to a spectrogram, a cochleagram is a T-F representation produced by filtering a signal using an auditory filterbank and then windowing each filter response into time frames (Wang and Brown, 2006, pp. 15–19). Aside from the target cochleagram, the same-talker masker condition is shown in the left column, the same-sex masker condition in the middle column, and the different-sex masker condition in the right column. The second row shows the masker cochleagrams in the three conditions, the third row shows the corresponding ideal binary masks where 1 is indicated by white and 0 by black, the fourth row shows the corresponding cochleagrams of the mixtures, and the bottom row shows the ITFS processed mixtures. There is general similarity among the three ideal masks, reflecting the fact that the same target utterance is used and all the masker utterances correspond to the same sentence. However, the ideal masks are different, and the difference is quite noticeable between the same-talker and different-sex talker conditions. For example, the right ideal mask has extended white regions at the bottom which are missing from the left ideal mask, because the first harmonic of the female masker occurs in a higher frequency range than the first harmonic of the male masker.

A visual inspection reveals a resemblance between the ITFS-processed stimuli in the bottom row of Fig. 1 and the original target signal in the top row. Because the ITFS technique removes energy from the masking speech without sig-

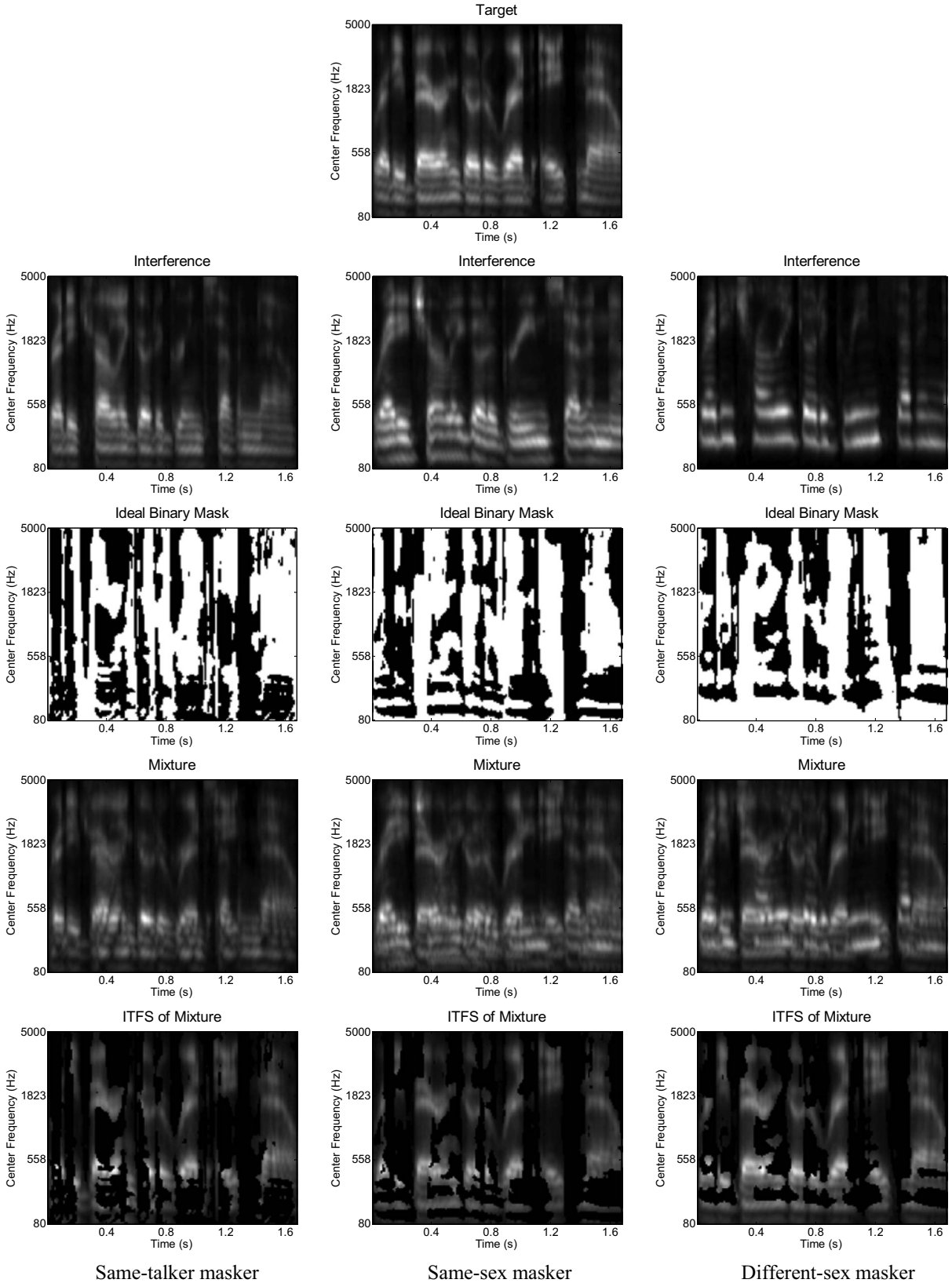

FIG. 1. ITFS illustration for mixtures of two utterances of the same talker, same-sex talkers, and different-sex talkers. Top row: Cochleagram of a target male utterance ("Ready Baron go to red eight now"). The figure displays the responses of the 128-channel gammatone filterbank, where the response energy at each T-F unit is raised to the $\frac{1}{4}$ power for better display. Second row: Cochleagrams of an interfering utterance ("Ready Charlie go to green seven now") spoken by the same talker (left), a same-sex talker (middle), and a different-sex talker (right). Third row: Corresponding ideal binary masks with 0 dB LC, where white pixels indicate 1 and black pixels indicate 0. Fourth row: Corresponding cochleagrams of the mixtures. Bottom row: Corresponding ITFS processed mixtures.

nificantly distorting the spectro-temporal pattern of the original target speech, it often produces large improvements in intelligibility when it is applied to a multitalker acoustic mixture. Brungart *et al.* (2006) reported that ITFS processing produced an intelligibility improvement that was approximately equivalent to a 22–25 dB decrease in the SRT of a multitalker speech signal.

Since ITFS processing does not divide the mixture energy within a T-F unit into target and masker portions but rather makes a binary, all-or-none decision on the mixture energy, it would be interesting to know whether the noise energy in a retained unit hinders speech intelligibility. Drullman (1995) examined a similar question by comparing speech intelligibility in two conditions using an auditory filterbank. In the first condition, noise was kept in the mixture where the noise level in a filter response was lower than the speech level; in the second condition, the noise was removed from the mixture. He found that removing noise that is below the speech level has no effect on intelligibility. On the other hand, a different comparison showed that substituting noise for speech that is below the noise level elevates the SRT by 2 dB (Drullman, 1995). This suggests that there is some useful speech information in T-F units where local SNR is negative, and this is consistent with the observation that a performance plateau with nearly perfect intelligibility centers at the LC value of −6 dB rather than the 0 dB LC (Brungart *et al.*, 2006)—the former LC retains more T-F units where local SNR is between 0 and −6 dB.

Although ideal binary masking techniques have been used for some time in computational auditory scene analysis, they have only recently begun to be applied as a psychophysical tool for measuring human auditory perception. Anzalone *et al.* (2006) tested the effects of processing with a related version of ideal binary mask, defined not in terms of a comparison between target energy and interference energy but a comparison between target energy and a predefined threshold. More specifically, a mask value of 1 was applied if and only if the corresponding target level exceeded a fixed threshold. They used mixtures of speech and speech-shaped noise, and documented the results in terms of SRT. The results show that ideal binary masking leads to substantial SRT reductions: more than 7 dB for normal-hearing listeners and more than 9 dB for hearing impaired listeners. Another intelligibility study by Li and Loizou (2007) used the ideal binary mask to generate "glimpses," or T-F regions with stronger target energy, to study several factors that impact glimpsing of speech in a mixture signal. Their results show that glimpses in the low- to mid-frequency range (up to 3 kHz) containing the first and the second formant of speech are particularly important for speech perception. They also showed that high intelligibility does not require perception of the entire utterance: glimpsing information in a majority of time frames (60%) seems to be sufficient for most intelligibility tasks.

III. EXPERIMENT 1: EFFECTS OF VOICE CHARACTERISTICS ON MULTITALKER LISTENING WITH ITFS

As stated earlier, differences in voice characteristics of competing talkers, in particular, the differences that exist between male and female voices, can increase speech perception performance (relative to the baseline case where the same talker is used for both the target and masking phrases) by as much as an equivalent increase of 7–11 dB in the SNR of the target speech (Festen and Plomp, 1990). In part, this improvement in performance occurs because reduced spectro-temporal overlap between the different-sex talkers allows the listener to "glimpse" a larger portion of the target speech. However, many other non-energetic factors also may contribute to this performance difference. In order to examine the influence of target and interferer similarity on the energetic component of speech-on-speech masking, an experiment was conducted that used the ITFS approach to compare multitalker listening performance with three levels of similarity between target and masking voices: a same-talker condition, where the target and interfering phrases were spoken by the same talker; a same-sex condition, where target and masking phrases were spoken by different talkers of the same sex; and a different-sex condition, where interfering phrases were spoken by talkers who were of the opposite sex of the target talker.

A. Methods

1. Stimuli

As in the previous study of Brungart *et al.* (2006), the speech stimuli used in this experiment were drawn from the publicly available coordinate response measure (CRM) speech corpus for multitalker communications research (Bolia *et al.*, 2000). The CRM corpus is based on a speech intelligibility test first developed by Moore (1981). The corpus contains phrases of the form "Ready (call sign) go to (color) (number) now." There are eight possible call signs ("Arrow," "Baron," "Charlie," "Eagle," "Hopper," "Laker," "Ringo," and "Tiger"), four possible colors ("blue," "green," "red," and "white"), and eight possible numbers (1–8). An example utterance is "Ready Baron go to blue five now." Eight talkers, four male and four female, were used to record each of the 256 possible phrases, resulting in a total of 2048 phrases in the corpus.

For each trial in the experiment, the target signal was a CRM phrase randomly selected from all the phrases containing the target call sign "Baron." The interference consisted of one, two, or three different phrases randomly selected from the CRM corpus that were spoken by the same talker, a different talker of the same sex, or a different talker of the opposite sex, depending on the particular condition of the experiment. Interfering phrases contained call-signs, color coordinates, and number coordinates that were different from the target phrase and different from each other. Each of the interfering phrases was scaled to have the same overall rms power as the target phrase, and then all the interfering

phrases were summed together to produce the overall interference used for ITFS. The target phrase and the interference were added to form the mixture signal.

Note the distinction between the target-to-masker ratio (TMR) and SNR, the former referring to the ratio of the target speech level to the level of each interfering talker and the latter to the ratio of the target talker level to the overall interference level. Hence, TMR was set to 0 dB in this experiment while SNR could be 0 dB or negative depending on how many competing talkers were included in the interference.

2. Listeners

Nine paid subjects participated in the experiment. The listeners all had normal hearing and their ages ranged from 18 to 54 years. Most had participated in previous auditory experiments, and all were familiarized with the CRM corpus and the experimental task prior to conducting this experiment.

3. ITFS segregation processing

Given a target signal, an interference signal, and a LC value, an ideal binary mask was constructed and used to resynthesize the mixture to generate a single ITFS stimulus (see Brungart *et al.*, 2006). Specifically, an input signal was first decomposed using a bank of 128 fourth-order gammatone filters with overlapping passbands (Patterson *et al.*, 1988) and with center frequencies ranging from 80 to 5000 Hz. Each filter response was further divided into 20 ms time frames with 10 ms overlap. Hence, the input signal was transformed into a matrix of T-F units. Within each T-F unit, the local SNR was calculated. If the local SNR was greater than or equal to the LC value, the ideal binary mask was assigned the value of 1 for this T-F unit and the mixture signal within the unit was included in the ITFS signal; if the local SNR was less than the LC value, the binary mask was assigned the value of 0 for the unit and the mixture signal within the unit was excluded from the ITFS signal. For further details on resynthesis from a binary mask, see Brungart *et al.* (2006) and Wang and Brown (2006, pp. 23–25).

Fourteen LC conditions were tested in the experiment, including values ranging from −48 to +30 dB. The LC values were chosen so that more values were tested in the range of 12–24 dB where speech intelligibility is known to drop sharply with increasing LC (Brungart *et al.*, 2006). An unsegregated control condition was also included, where the stimulus was generated by applying ITFS processing to a mixture signal with an all-1 mask. Hence, the control condition amounted to presenting the mixture to a listener after equalizing for any possible distortion that might be introduced by ITFS processing. In all cases, the overall presentation level of the stimulus was kept approximately constant (roughly 65 dB sound pressure level) across all LC values by scaling the maximum value of the ITFS-processed mixture waveform to a fixed level prior to presentation of the stimulus.

4. Procedure

During the experiment, a listener was seated at a control computer in a quiet listening room. The stimuli were generated by a sound card in the control computer (Soundblaster Audigy) and presented to the listener diotically over headphones (Sennheiser HD-520). On each trial, the listener was instructed to use the mouse to select the colored digit corresponding to the color and number of the target phrase (containing the call sign "Baron") on an eight-column, four-row array of colored digits corresponding to the response set of the CRM task.

The experiment was divided into three sub-experiments, with each sub-experiment examining performance with two, three, or four competing talkers over a specific range of LC values. Because some LC values were repeated across sub-experiments, this resulted in an uneven distribution of data collection across the 15 LC conditions tested in the experiment, with some additional trials collected at LC values that overlapped across the sub-experiments. The stimuli for each listener were selected randomly prior to the start of the experiment, processed off-line, and stored on a personal computer (PC) for later presentation to the listeners.

Seven subjects participated in a complete set of trials across the three sub-experiments, resulting in a total of 4500 trials, divided into blocks of 50 trials. The different numbers of talkers and the different characteristics of the interfering talker(s) were evenly distributed across all the trials. Thus, for these subjects, there were 500 trials for each of the nine configurations (2-, 3-, and 4-talkers by same-talker, same-sex, and different-sex configurations). Two additional subjects completed only part of the experiment (3000 trials and 600 trials, respectively). Data from these subjects were included in the overall data analysis, but they were excluded from analyses that required the calculation of separate thresholds for the individual subjects.

B. Results and discussion

Figure 2 shows the percentage of correct color and number identifications as a function of the LC value used to generate the ITFS stimulus for all the configurations in the experiment. The figure is divided into three panels to separate the results for 2, 3, and 4 simultaneous talkers. Within each panel, the three curves show performance for three different levels of similarity between the target and the interfering voices: same-talker, same-sex, and different-sex. Each data curve was generated by fitting with two logistic functions: one for LC values greater than 0 dB, and one for LC values less than 0 dB (Cavallini, 1993). The data in the curves were averaged across listeners, and the error bars in the figure represent the 95% confidence interval for each data point.

The leftmost points of each curve in Fig. 2 show the results from the unsegregated control condition where all of the T-F units in the original mixture were retained in the processed speech. In the same-talker conditions, the listeners correctly identified both the color and number in the target phrase in 8% of the trials in the 4-talker condition, 23% in the 3-talker condition, and 45% in the 2-talker condition.

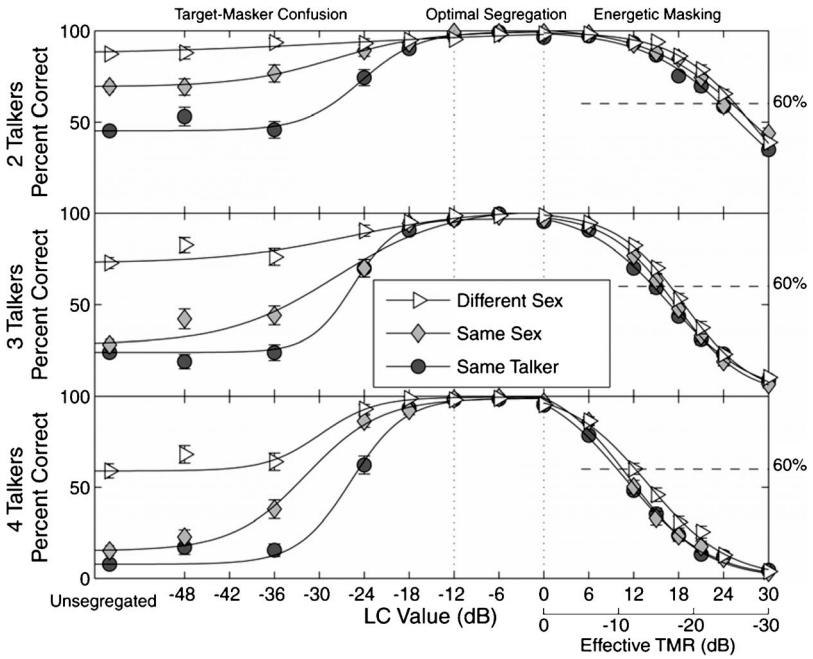

FIG. 2. Percentage of correct color and number identifications in Experiment 1 as a function of LC. The top panel shows results for the 2-talker conditions, the middle panel shows results for the 3-talker conditions, and the bottom panel shows results for the 4-talker conditions. The error bars represent 95% confidence intervals (±1.96 standard errors, calculated from the pooled response data across all listeners) in each condition. Horizontal dashed lines indicate the 60% threshold levels of performance at positive LC values.

This level of performance is comparable to that achieved in the unsegregated conditions of our previous study examining the effect of ITFS processing on same-talker speech (Brungart *et al.*, 2006). Switching from same-talker interfering voices to same-sex interfering voices improved performance substantially in the 2-talker condition but produced a relatively modest improvement in performance in the 3- and 4-talker conditions. Switching from same-sex interfering voices to different-sex interfering voices, however, produced substantial performance improvements in all the configurations tested. These results are consistent with those that have been obtained in other experiments that examined the effect of target-interferer similarity on 2-, 3-, and 4-talker listening at a 0 dB TMR (Brungart *et al.*, 2001).

1. ITFS regions

As was the case in the earlier ITFS study by Brungart *et al.* (2006), the experimental results show that there are three distinct regions of intelligibility performance: an "optimal segregation" region with $0\ \text{dB} \geq \text{LC} \geq -12$ dB, a "target-masker confusion" region with $\text{LC} < -12$ dB, and an "energetic masking" region with $\text{LC} > 0$ dB. In the optimal segregation region, performance was nearly perfect in all conditions tested. In this region, the application of the ITFS processing essentially removed all the audible elements of the masking signals from the stimuli, thus eliminating informational masking related to target-masker confusion. At the same time, the processing preserved enough of the target information to allow near perfect color-number intelligibility in the CRM task. It is clear from these results that, even in extreme cases where an acoustic mixture contains four simultaneous voices spoken by the same talker, there are enough T-F units dominated by the target voice to allow a listener to almost perfectly extract the meaning of the target message.

When the LC value was reduced below −12 dB, performance began to steadily decrease. In this target-masker confusion region, the LC value was increasingly negative, which means that the stimulus included an increasing number of T-F units where the masker was more powerful than the target speech. This did not remove any acoustic information from the target speech signal, per se, although it is possible that in some cases the addition of a relatively high-level masker-dominated T-F unit may have obscured some portions of the target signal due to non-simultaneous masking, either upward or downward in frequency or forward or backward in time. The more profound impact that the addition of these masker-dominated T-F units had on performance was an increased probability of incorrectly grouping the interferer-dominated T-F units retained in the ITFS stimulus with the T-F units in the target speech. These confusions resulted in a dramatic increase in the number of trials where the listener incorrectly responded with the number and color keywords contained in the masking phrase. In our earlier study, we showed that, in the 2-talker condition, nearly 100% of the incorrect responses in the target-masker confusion region matched the color and number keywords spoken by the interfering talker (Brungart *et al.*, 2006). In this experiment, this trend was seen across all the target-masker similarity conditions with 2-, 3-, or 4-talkers, with more than 96% of the incorrect color responses and more than 94% of the incorrect number responses in this region matching the keywords spoken by one of the interfering talkers.

The breakpoint in the LC-performance function at −12 dB represents the point where the interferer-dominated T-F units retained in the ITFS stimulus began to resemble an intelligible masking talker. At LC values greater than −12 dB, the interferer-dominated T-F units introduced into the mixture occurred in T-F regions where the target speech also has relatively significant energy, reducing the probability that those units will be interpreted as an additional masking voice. When the LC value was less than −12 dB, the ITFS stimulus included T-F units where the target speech has relatively very weak energy, and the inclusion of these T-F

TABLE I. Threshold effective TMR values for 60% performance in Experiment 1. These values are obtained from the secondary scale at the lower right of Fig. 2.

	Interferer(s) voice characteristics					
	Same-talker	Same-sex		Different-sex		Mean
2 Talkers	−23.68 dB	−25.14 dB	Δ=−1.46 dB	−25.51 dB	Δ=−0.37 dB	−24.78 dB
3 Talkers	−13.73 dB	−15.14 dB	Δ=−1.41 dB	−16.44 dB	Δ=−1.30 dB	−15.10 dB
4 Talkers	−8.88 dB	−9.58 dB	Δ=−0.70 dB	−10.68 dB	Δ=−1.10 dB	−9.72 dB
Δ Mean			Δ=−1.19 dB		Δ=−0.93 dB	

units appeared to produce a substantial amount of informational masking, particularly in the same-talker and same-sex conditions where the masking voices were qualitatively similar to the target voice. In the limit, where the LC value was set to −48 dB, the stimulus essentially contained both the target and masking speech signals in their entirety, and performance asymptoted at the same level as the unsegregated control condition.

For the purposes of this study, the most informative region of Fig. 2 is the energetic masking region, where the LC value was systematically increased above 0 dB. In this region, overall performance decreased with increasing LC value because an increasingly large proportion of the target speech signal failed to meet the threshold SNR value and thus was eliminated from the stimulus. To the extent that the ITFS methodology captures the temporal and spectral resolution of the auditory system (see Brungart *et al.*, 2006), one would expect performance in this region to be directly related to the amount of energetic masking caused by the spectro-temporal overlap between the target and masking signals. From the results in the figure, it is clear that target-masker similarity had a much smaller impact on performance in the energetic masking region than it did in the target-masker confusion region, where performance was limited by the ability to separate the T-F units associated with target and masker voices rather than by the absence of target-dominated T-F units due to the spectro-temporal overlap in the competing signals. However, it is clear that there was a small but consistent advantage for the different-sex speech in the energetic masking region of Fig. 2.

In order to quantify the magnitude of this advantage, it is helpful to take advantage of the close relationship that exists between LC value and TMR in the energetic masking region (Brungart *et al.*, 2006). In that region, each 1 dB increase in LC value removes the same T-F units from the target speech that would be lost due to the increased spectro-temporal overlap caused by a 1 dB decrease in the SNR of the acoustic mixture, or equivalently a 1 dB decrease in the TMR. Thus, for example, an acoustic mixture with a TMR value of 0 dB that is ITFS processed with a LC value of 6 dB retains exactly the same set of target-dominated T-F units as an acoustic mixture with a TMR value of −6 dB processed with a LC value of 0 dB. Therefore, each data point in the energetic masking region can be viewed as an estimate of optimally segregated ITFS performance for an acoustic mixture with an effective TMR value equal to the negative of the LC value plotted in the figure. These effective TMR values are shown in the secondary scale at the lower right of Fig. 2.

In order to calculate the overall thresholds in each condition, the curves in each panel of Fig. 1 were fitted separately for each of the seven listeners who completed the entire experiment and used to calculate individual thresholds for 60% correct performance for each combination of target-interferer similarity and number of interfering talkers. The resulting mean threshold values are presented in Table I. If the threshold values for each target-masker condition are averaged across the number of talkers, it is apparent that changing from same-talker interfering voices to same-sex interfering voices produced, on average, only a 1.19 dB decrease in the 60% threshold TMR value, and changing from same-sex interfering voices to different-sex interfering voices only produced an additional 0.93 dB decrease in the 60% threshold. In order to evaluate the statistical significance of these differences, the individual 60% thresholds were subjected to a two-factor, within-subject analysis of variance (ANOVA). The results of this ANOVA show that the main effects of target-interferer similarity $[F(2,12)=883.357]$ and number of competing talkers $[F(2,12)=28.418]$ were both significant at the $p<0.001$ level, but that their interaction was not significant $[F(4,24)=0.395,\ p=0.740]$. Thus, on the basis of these results, it seems that target-interferer similarity does have a significant impact on energetic masking in multitalker listening, but that it can account for no more than a 2–3 dB change in the energetic masking effectiveness of an interfering speech signal.

These values can be compared to the results of previous experiments that have used different methods to estimate the differences in the energetic masking efficiency of same- and different-sex voices. As noted in Sec. I, Festen and Plomp (1990) used speech-shaped noise to estimate the energetic masking component of same- and different-sex talkers, and found a difference of less than 2 dB across the two types of masking voices, versus a 7–11 dB difference for normal speech maskers. Thus the results obtained with a speech-shaped noise masker are roughly comparable to those obtained with the ITFS method, despite the substantial differences in methodology.

2. Comparison to statistical analyses of ITFS-processed speech

From the psychoacoustic results shown in Fig. 2, it is clear that there is a small but systematic increase in energetic masking when an interfering voice is made more similar to the target talker, and a large increase in energetic masking when the number of interfering talkers increases. Certainly

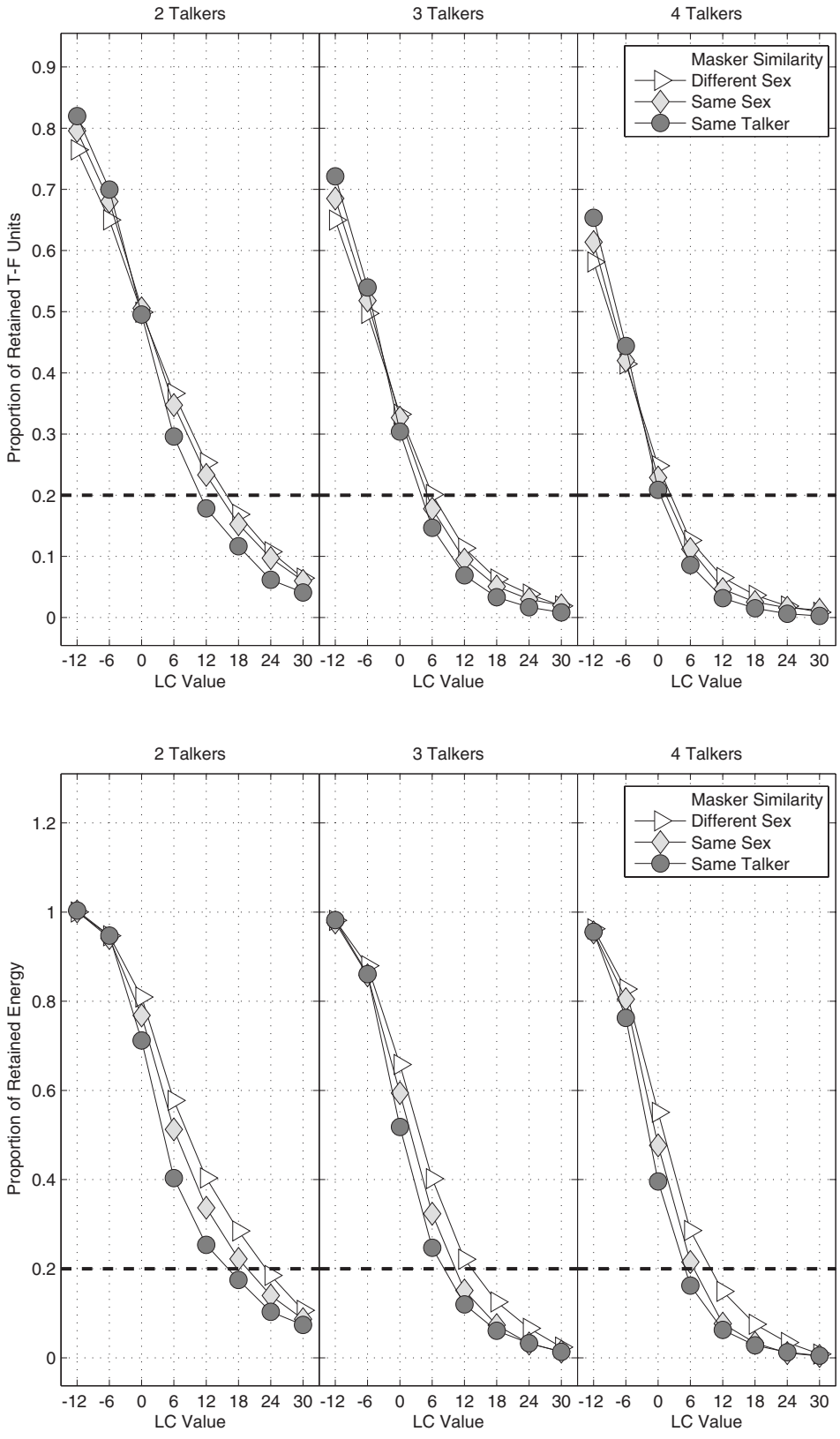

FIG. 3. Acoustic analysis of ITFS-processed stimuli in each masking condition. The top panel shows the proportion of T-F units retained in each stimulus as a function of the LC value. The bottom panel shows the proportion of total energy retained in the target stimulus as a function of the LC value. See text for details.

an interesting question related to these results is the extent to which these psychometric results correlate with the acoustical properties of the target signal that was retained in the ITFS stimulus in each condition. Figure 3 shows two analytical measures of the proportion of the target signal retained after the ITFS processing at each LC value for each masker condition tested in the experiment. The top panel shows the proportion of retained T-F units in each condition. These values were obtained by randomly generating roughly 400 stimuli for each data point using the same procedure used in the psychophysical experiment, and simply counting the number of T-F units retained after the ITFS processing. The

TABLE II. Threshold effective TMR values for retention of 20% of the total T-F units within each type of ITFS-processed stimulus.

	Interferer(s) voice characteristics				
	Same-talker	Same-sex		Different-sex	
2 Talkers	−11.21 dB	−14.29 dB	Δ=−3.08 dB	−15.44 dB	Δ=−1.15 dB
3 Talkers	−3.81 dB	−5.26 dB	Δ=−1.45 dB	−6.22 dB	Δ=−0.96 dB
4 Talkers	−0.45 dB	−1.32 dB	Δ=−0.87 dB	−2.18 dB	Δ=−0.86 dB
Δ Mean			Δ=−1.80 dB		Δ=−0.99 dB

bottom panel shows the proportion of *energy* retained in the stimulus, calculated simply by applying the ideal binary mask for the trial to the target *only* (not the mixture) and determining how much of the target energy was contained in the T-F units that were retained in the ITFS stimulus.

The overall trends of the retained unit and retained energy curves in Fig. 3 are similar to those for the psychoacoustic data in the energetic masking region of Fig. 2, with an increase in the retained portion of the target signal as the similarity between the target and masker decreased. This can be clearly seen in Tables II and III, which show the threshold effective TMR values for 20% retained units and 20% retained energy, respectively, using the same procedure used to calculate the 60% correct individual thresholds in Table I. The overall effects of target-masker similarity were most similar to the psychoacoustic results in the retained-bin metric, where there was a 1.8 dB shift in the threshold for 20% retained units between the same-talker and same-sex conditions (compared to 1.2 dB for the psychoacoustic data), and an additional 1.0 dB difference in the threshold for 20% retained units between the same-sex and different-sex conditions (compared to 0.9 dB for the psychoacoustic data). The shifts for 20% retained energy thresholds shown in Table III were also similar in direction but slightly larger in magnitude than the corresponding shifts in the psychoacoustics results, with a 2.5 dB shift between the same-talker and same-sex conditions and a 3.1 dB shift between the same-sex and different-sex conditions.

Another notable feature shown in the upper panel of Fig. 3 is a reversal in the ordering of the proportion of retained units for the different target-masker similarity conditions when the LC value was less than 0 dB. Indeed, at negative LC values, there were actually more retained units in the same-talker conditions than in the different-sex conditions. This result reflects a fundamental relationship between spectro-temporal overlap and the proportion of retained units in the ideal binary mask paradigm. When the LC value is positive, there is a tendency to eliminate target units in T-F regions where both the target and the masker contain energy, so there is generally a reduction in retained units when the spectro-temporal overlap in the target and masking signals increases. However, when the LC value is negative, this trend is reversed, and there is a general tendency to retain T-F units where the target and masker overlap. Thus, at negative LC values, the proportion of retained target units actually *increases* as the target and masker become more similar. As an illustration of this principle, consider two extreme cases, one where the target and masker have exactly the same energy distribution, and one where they are completely non-overlapping (each occupying 50% of the T-F units). In the completely overlapping case, the proportion of retained units will go from 0% to 100% as the LC value changes from positive to negative, because the local SNR value is exactly 0 dB in every T-F unit. In contrast, in the non-overlapping case, the proportion of retained units remains at 50% independent of the LC value.

In the retained energy proportions shown in the lower panel of Fig. 3, there is no evidence of a reversal in the relative ordering of the different target-masker similarity conditions at negative LC values. This seems to be at least partially related to a ceiling effect in the results. Even in the 4-talker conditions, most of the energy was already retained in the stimulus when the LC value was −6 dB.

If the top and bottom panels are carefully compared across the three different columns of Fig. 3, it is apparent that the ratio of the percentage of retained energy to the percentage of retained units systematically increased as the number of competing talkers increased. This result is most likely related to the changes in the distribution of the energy in the masker. As the number of competing talkers in the stimulus increased, the spectro-temporal distribution of energy in the maskers became more uniform, which reduced the probability of retaining T-F units that contained only a

TABLE III. Threshold effective TMR values for retention of 20% of the total target energy within each type of ITFS-processed stimulus.

	Interferer(s) voice characteristics				
	Same-talker	Same-sex		Different-sex	
2 Talkers	−15.01 dB	−18.76 dB	Δ=−3.75 dB	−21.93 dB	Δ=−3.17 dB
3 Talkers	−7.80 dB	−9.96 dB	Δ=−2.16 dB	−13.14 dB	Δ=−3.18 dB
4 Talkers	−4.68 dB	−6.36 dB	Δ=−1.68 dB	−9.39 dB	Δ=−3.03 dB
Δ Mean			Δ=−2.53 dB		Δ=−3.13 dB

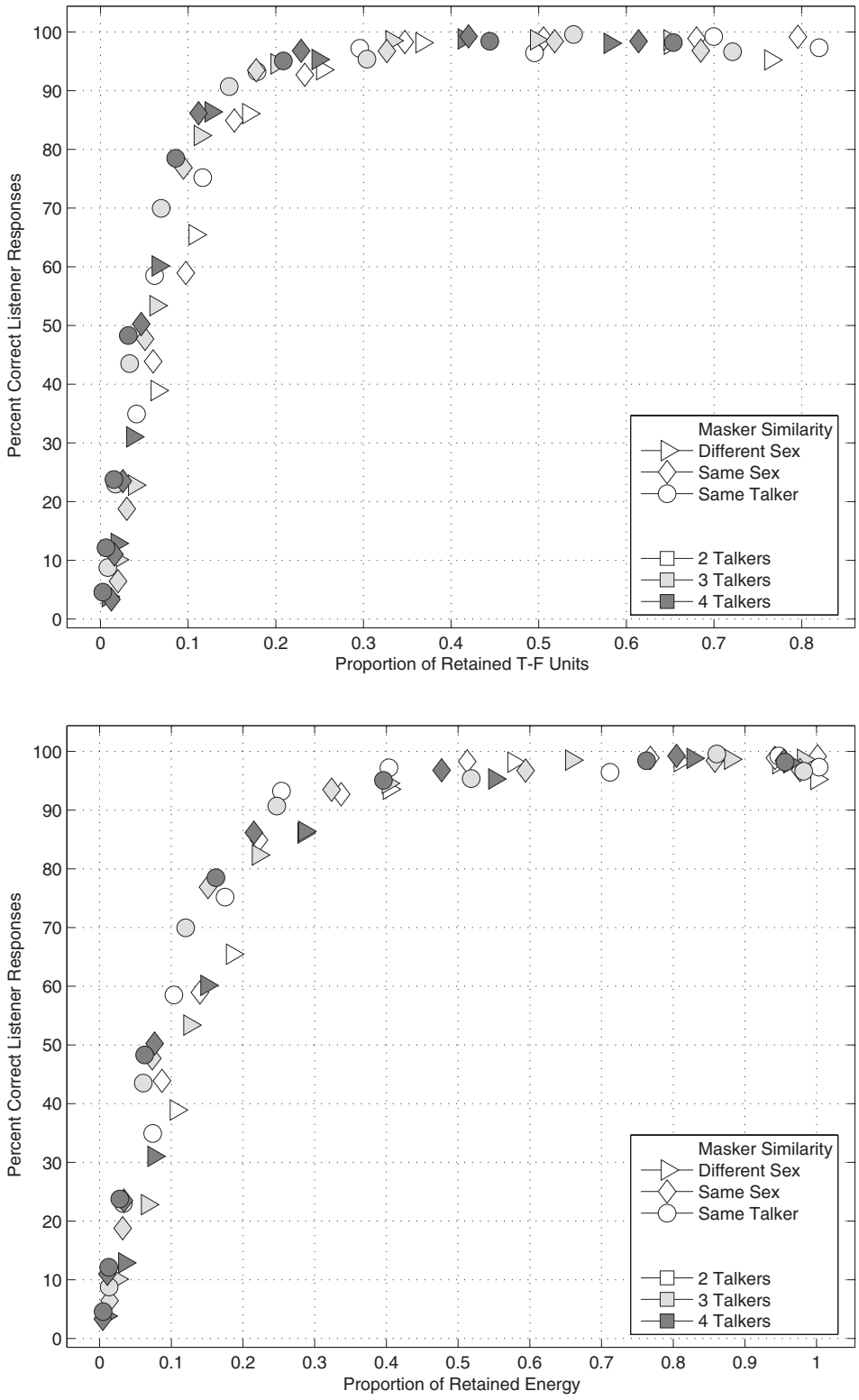

FIG. 4. Scatter plot showing the percentage of correct color and number responses in Experiment 1 as a function of the average proportion of retained T-F units (top panel) and average proportion of retained target energy (bottom panel), in each stimulus condition.

small amount of target energy. This had the net effect of increasing the average amount of target energy per unit in the ITFS stimulus.

The proportions of retained units and retained energy also appeared to do a good job of predicting the overall intelligibility of an ITFS stimulus across different numbers of talkers and different LC values in Experiment 1. This can be seen from Fig. 4, which shows scatter plots of the percentage of correct listener responses in each masker configuration as a function of the proportion of retained units (upper plot) and the proportion of retained target energy (lower plot). In general, the relatively tight distributions of data points in these two scatter plots show that listener performance was similar across all the stimulus conditions that resulted in a similar proportion of retained units or a similar proportion of retained target energy. However, there are two notable trends in the data. In the retained unit plot (upper panel), there was a general tendency for performance in the 2-talker condition

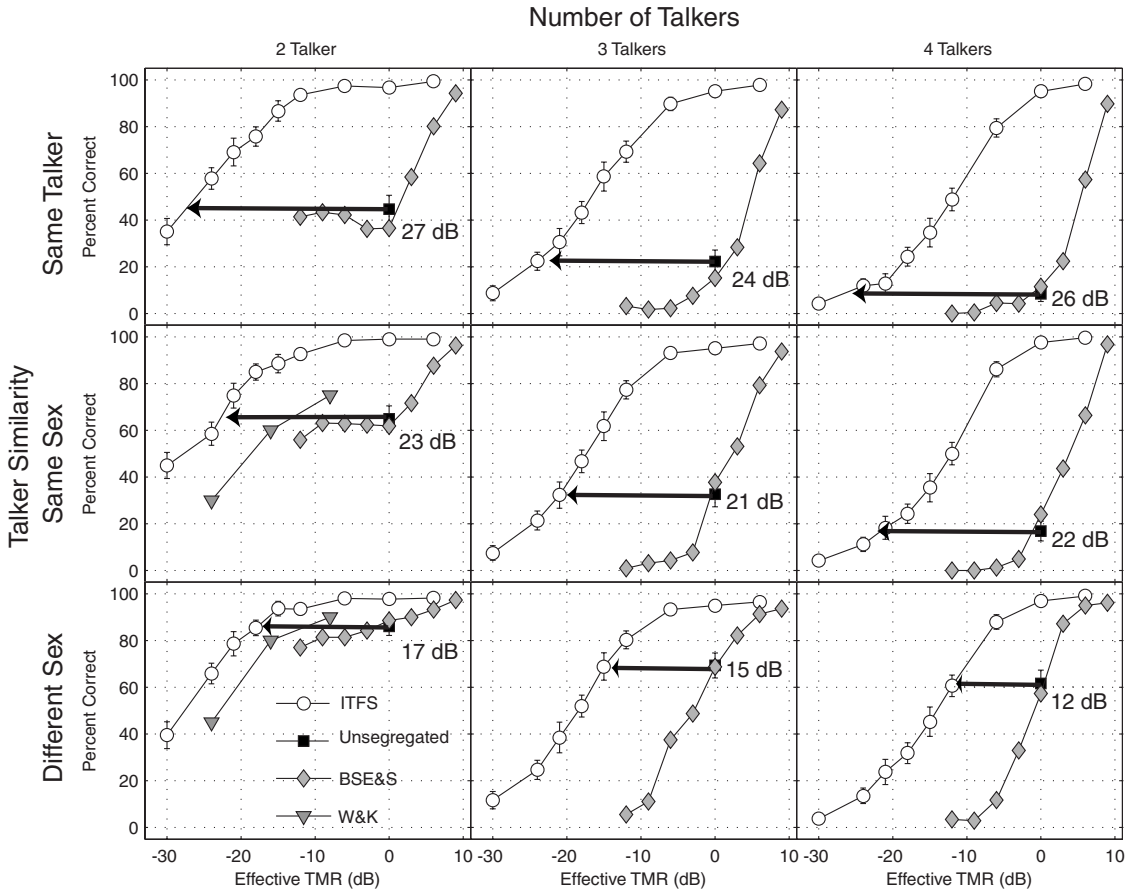

FIG. 5. Comparison of performance with ITFS-processed and unsegregated stimuli as a function of the effective TMR value of the stimulus. The open circles, which are re-plotted from Fig. 2, show performance in the ITFS-processed conditions of the experiment as a function of the effective TMR value of each data point. The filled squares show performance in the unsegregated condition of the experiment. The filled diamonds and filled triangles show performance from two earlier studies that measured multitalker listening performance with unsegregated CRM stimuli as a function of TMR (Brungart *et al.*, 2001, and Wightman and Kistler, 2005, respectively). The black arrows show in decibels the release from masking obtained in the ITFS condition relative to the unprocessed condition at a TMR value of 0 dB.

(white symbols) to fall slightly below performance in the same-talker and same-sex conditions (light and dark gray symbols) with the same proportion of retained units, especially in the range from 0.1 to 0.2 retained units. Similarly, in the retained energy plot (lower panel), there was a general tendency for performance in the different-sex condition (triangles) to fall slightly below performance in the same-sex and same-talker conditions (diamonds and circles) with the same proportion of retained energy.

3. Evaluation of non-energetic masking effects

By comparing the results in Fig. 2 to previous psychophysical results that have measured the effect of TMR on speech perception with the CRM corpus, it is possible to calculate a rough estimate of the amount of non-energetic masking that occurs in each condition. The open circles in Fig. 5 redraw a subset of the data from Fig. 2 in a nine-panel figure, with each row of the figure corresponding to a different level of target-masker voice similarity and each column corresponding to a different number of talkers in the stimulus. Within each panel, the black square at a TMR value of 0 dB shows the level of performance achieved in the unsegregated control condition of this experiment. For comparison, the gray diamonds in the figure show unsegregated performance as a function of TMR from a study that used the same CRM materials used in this experiment (Brungart *et al.*, 2001). These results are nearly identical to the unsegregated results collected at a TMR value of 0 dB in the present experiment. In the 2-talker same-sex and different-sex conditions, shaded triangles have been used to plot the results of another study (Wightman and Kistler, 2005) that examined performance in monaural presentations of the CRM stimuli to adult listeners as a function of TMR value at lower TMR values than those used in the Brungart *et al.* (2001) study. This experiment used only two of the eight talkers in the CRM corpus, and this resulted in generally higher performance levels that are probably not directly comparable to the present study, but the data do provide some insights into the psychometric function for 2-talker CRM listening tasks at very low TMR values.

In all cases, it is clear that the ITFS processing resulted in a substantial improvement in performance over the unsegregated condition conducted at the same effective TMR value. The black arrows show the decrease in effective TMR value required to bring performance with the ITFS-processed mixture down to the same level achieved with the original (unprocessed) mixture at a TMR value of 0 dB. The size of this shift, which is indicated by the number shown to the

right of the arrow, can be viewed as a crude quantitative estimate of the amount of non-energetic masking in each listening condition. Across the nine conditions tested in this experiment, the advantage of the ITFS processing ranged from 12 dB in the 4-talker, different-sex condition to 27 dB in the 2-talker, same-talker condition. These results suggest that, in contrast to the amount of energetic masking, the amount of non-energetic masking is influenced more by the target-masker similarity than by the number of competing talkers in the stimulus. Across all conditions, the amount of non-energetic masking was 6–10 dB higher in the same-sex condition than in the different-sex condition, and an additional 3–4 dB higher in the same-talker condition than in the same-sex condition.

Although the main focus of Experiment 1 was an examination of the effects of target-masker similarity on energetic and non-energetic masking for a fixed number of talkers, the relatively modest variations in non-energetic masking that occurred when the number of interfering talkers was increased at a fixed level of target-masker similarity are also noteworthy. These effects can be seen by looking across the arrows from left to right in each row of Fig. 5. In the same-talker and same-sex conditions, the amount of non-energetic masking was almost constant across the 2, 3, and 4-talker conditions, varying only from 24 to 27 dB in the same-talker masker condition and from 21–23 dB in the same-sex masker condition. In the different-sex masking condition, the amount of non-energetic masking decreased more substantially (from 12 to 17 dB) as the number of interfering talkers increased from 2–4, but this change was still quite modest in comparison to the much larger 9–14 dB changes seen with variations in the target-masker similarity of the stimulus.

Qualitatively, these systematic changes in non-energetic masking with the number of competing talkers appear to be smaller than those reported by Freyman *et al.* (2004), who reported roughly a 5 dB decrease in non-energetic masking as the number of same-sex interfering talkers increased from 2 to 3, and by Simpson and Cooke (2005), who did not use a decibel-based measure but reported an apparent *increase* in non-energetic masking as the number of interfering talkers increased from 1 to 3. However, direct comparisons of non-energetic masking across different measurement methodologies and, in particular, across different speech perception tests should only be made with caution: many factors will influence the amount of non-energetic masking, including the underlying confusability of the target and masking stimuli (relatively high with the CRM stimuli) and the underlying sensitivity of the speech perception test to energetic masking (relatively low with the CRM stimuli). The method for estimating non-energetic masking shown in Fig. 5 is a useful method because it provides a consistent means for comparing informational masking effects across a broad range of different stimulus types. However, it would be a mistake to assume that all speech perception tests will produce the same results obtained with the CRM stimuli. Furthermore, it is difficult to extrapolate the results shown in Fig. 5 to those that would be obtained with a larger number of competing talkers. Experiment 2 was conducted to more fully examine how energetic and non-energetic masking varied as a function of the number of competing talkers in the CRM test.

IV. EXPERIMENT 2: EFFECTS OF NUMBER OF COMPETING TALKERS WITH ITFS

One interesting aspect of the results from Experiment 1 is that none of the conditions tested indicates that any significant amount of energetic masking occurs in a CRM task with a TMR of 0 dB. Even in the most difficult condition tested where the target CRM phrase was masked by three interfering phrases spoken by the same talker who produced the target speech, performance was near 100% when the LC value was set to 0 dB (as compared to about 12% correct in the "unsegregated" condition that included the effects of both informational and energetic masking). This raises the interesting question of how many simultaneous overlapping equal-level speech signals really are necessary to produce a considerable amount of energetic masking in the CRM listening task. In order to explore this question, a second experiment was conducted to examine performance as a function of the number of competing talkers when the LC value was fixed at 0 dB.

A. Methods

The procedures used in Experiment 2 are similar to those used in Experiment 1. For each trial, the target and interfering phrase(s) were randomly selected from the CRM corpus, scaled to have the same overall rms levels, and summed together to produce target, interferer, and mixture signals. The procedures outlined in Sec. III were then used to generate an ideal binary mask with LC set to 0 dB, and this binary mask was used to resynthesize an output signal that was stored off-line on a PC for later presentation to the listeners. The primary difference between Experiment 2 and Experiment 1 was the number of interfering talkers: in Experiment 2, the number of interfering talkers in each trial was randomly selected from 1 of 10 values ranging from 1 to 18 (1, 2, 4, 6, 8, 10, 12, 14, 16, and 18), with all of the interfering phrases spoken by the same talker (randomly selected on each trial) who spoke the target phrase.

Due to the large number of simultaneous voices for some cases, as well as the limiting variety of unique call-sign/color/number combinations in the CRM corpus, extra care was taken to ensure that the phrases presented within each trial differed in terms of call-sign, color, and number to the greatest extent possible. Previously, call signs, colors, and numbers could not repeat within the same trial. In this experiment, while the target call sign remained unique with "Baron," the interferer call signs, colors, and numbers could duplicate within each trial. The distribution of these conditions within each trial was assigned to minimize such duplication. The same nine listeners who participated in Experiment 1 also participated in Experiment 2, with each subject conducting 10 blocks of 50 trials.

B. Results and discussion

The overall results from Experiment 2 are shown by the shaded circles in Fig. 6. In this figure, each shaded circle

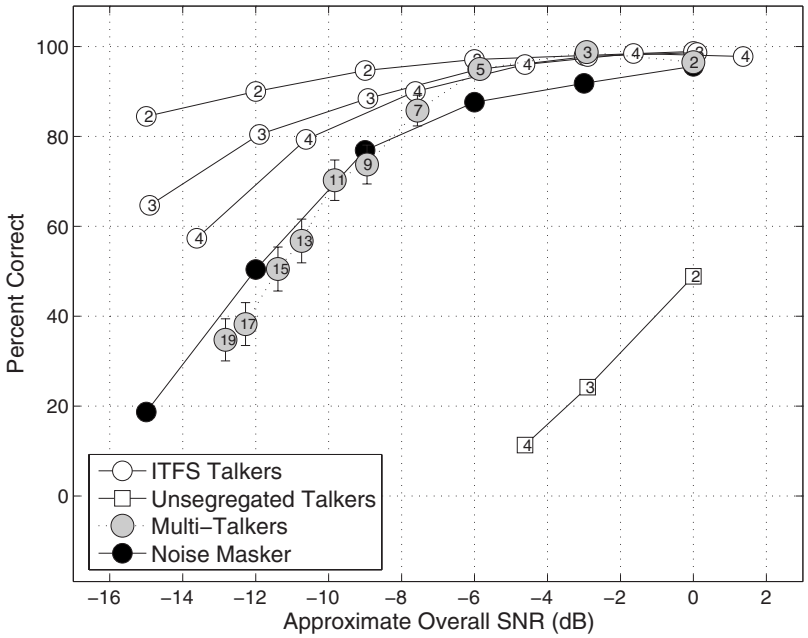

FIG. 6. Percentage of correct color and number identifications in Experiment 2 as a function of the overall SNR. The shaded circles show the performance for each of the ten different numbers of simultaneous talkers tested in Experiment 2, with the number of simultaneous talkers in each condition indicated in the center of each data point. In each case, the data have been plotted as a function of the mean overall SNR. The error bars show the 95% confidence intervals around each data point. For comparison purposes, the open and numbered circles re-plot the data obtained for 2-, 3-, or 4-talkers at different positive LC values to show performance as a function of the approximate overall SNR in a stimulus containing a fixed number of interfering talkers. The open squares show results from the unsegregated conditions with 2-, 3-, or 4-talkers and a TMR value of zero (plotted here as a function of SNR). The closed symbols show performance for a CRM stimulus with an ITFS-processed continuous noise masker (replotted from Brungart *et al.*, 2006).

represents a different number of competing talkers, as indicated by the number at the center of each data point. In each case, the data have been plotted as a function of the mean overall SNR of the original stimulus, calculated from the ratio of the total rms energy in the target phrase to the total rms energy in the interference of a multitalker mixture. For example, in the 3-talker case of Experiment 2 (shaded circles), the two interfering talkers were presented at the same level as the target speech, so the total mixture of interfering signals in that condition is roughly 3 dB more intense than the target and thus produces an overall SNR of approximately −3 dB. In the 19-talker condition, the 18 interfering talkers produce a combined rms interference energy level roughly 13 dB higher than the target phrase, so it is plotted at a SNR value of approximately −13 dB.

As in Experiment 1, these results show that the color and number identification performance is near 100% when the stimulus contained a small number of competing talkers (≤4). As the number of competing talkers is increased beyond 4, there was a steady decrease in performance. However, even in the worst condition tested, where the target speech signal was masked by 18 different competing speech signals each spoken at the same level by the same talker, the application of ITFS with 0 dB LC resulted in a performance level that was well above chance (35% versus 3% for chance performance). These results suggest that the difficulties listeners have in understanding signals containing a large number of competing talkers are largely due to informational, not energetic, masking effects. Put another way, it seems that even extremely dense multitalker babbles often contain enough glimpses of the individual talkers to make it theoretically possible to comprehend a single voice, but without some way to conclusively group the glimpses produced by a single talker together into a cohesive auditory image, the listener has no hope of successfully understanding any of the individual talkers in a complex multitalker environment.

The results from Experiment 2 also provide an opportunity to further explore the energetic masking efficiency of an interfering speech signal. A number of previous researchers have commented that masking signals containing more than one interfering talker generally produce lower intelligibility than masking signals containing a single interfering talker at the same overall SNR. This is believed to occur because the multiple competing speech signals "fill in the gaps" that listeners could ordinarily use to obtain a glimpse of the target speech (Simpson and Cooke, 2005; Freyman *et al.*, 2004; Bronkhorst and Plomp, 1992; Festen and Plomp, 1990; Miller, 1947). To this point, however, it has been difficult to accurately measure this effect because of the complications involved in distinguishing the increased energetic masking caused by filling in the gaps from any change in informational masking that might occur when more voices are added to the stimulus. The ITFS approach used in this study allows us to make a direct comparison of the effects of energetic masking as a function of the number of interfering talkers in the stimulus and the effective overall SNR of the signal.

The open circles in Fig. 6 re-plot the data obtained at the different positive LC values tested in Experiment 1 of Brungart *et al.* (2006) to show performance as a function of the approximate effective overall SNR in a stimulus containing a fixed number of interfering talkers (these data were collected with the same nine listeners used in this study; they are plotted here because the earlier study tested more LC values in the same-talker condition) For example, in the 4-talker condition of that experiment, the ITFS processing with a LC value of 0 dB produces results that roughly simulate the energetic masking that occurs from a 3-talker interfering speech signal presented at a TMR of 0 dB. Such an interfering signal, which contains three speech signals with the same rms power as the target speech, has a combined rms power that is approximately 4.8 dB higher than the target. Consequently, the 4-talker condition with a 0 dB LC value is plotted at an SNR value of −4.8 dB in Fig. 6. Similarly, the 4-talker data collected with a LC value of +3 dB corresponds to a 4-talker signal with a mixture SNR reduced by 3 and 0 dB LC, giving an overall SNR of approximately −7.8 dB.

Thus, the 4-talker data with a LC value of +3 dB from Experiment 1 of Brungart *et al.* (2006) is re-plotted in Fig. 6 at an SNR value of −7.8 dB. Similar procedures are used to plot performance from the 2-, 3-, and 4-talker conditions of that experiment as a function of overall effective SNR for all SNR values from 0 to −13 dB.

For comparison purposes, data are also shown for unsegregated speech maskers and for ITFS-processed noise maskers. The open squares in Fig. 6 re-plot the data from the unsegregated 2-, 3-, and 4-talker conditions of Brungart *et al.* (2006) as a function of stimulus SNR. Finally, the closed symbols in Fig. 6 show performance for the ITFS processed continuous noise masker from Experiment 2 of Brungart *et al.* (2006) as a function of the effective stimulus SNR. In theory, this type of speech-shaped noise masker is equivalent to a masking signal comprised of an infinite number of interfering talkers.

Comparing the different curves in Fig. 6, it is first apparent that performance in the unsegregated condition (open squares) dropped off much more dramatically with an increase in the number of competing talkers than any of the ITFS-processed conditions. Indeed, when the number of unsegregated talkers increases to 4, performance dropped to near chance level, even though the effective SNR value in that condition is greater than −5 dB. In comparison, performance in the ITFS conditions of Experiment 2 remained well above chance level even when there were 19 simultaneous talkers and the SNR was less than −12 dB. Thus one can conclude that non-energetic masking effects related to the confusability of the target and masking voices tend to dominate overall performance in CRM tasks involving more than one simultaneous talker.

Next, it is apparent that the effects of energetic masking indeed increase substantially when additional interfering voices are added to a stimulus in which the overall SNR is held constant—for instance, compare the percent correct values of the curves at the overall SNR of −10 dB. The 2-, 3-, and 4-talker lines show a systematic increase in energetic masking with the number of competing talkers, and the "multitalker" line from Experiment 2 shows considerably worse performance than the 4-talker line for all conditions containing more than four competing talkers. In the most extreme case tested, the 19-talker point from Experiment 2 shows performance dropping to roughly 35% correct responses when the overall SNR is −13 dB, compared to performance levels of approximately 60%, 75%, and 85% in the 4-, 3-, and 2-talker conditions, respectively, at the equivalent overall SNR. These results conclusively demonstrate that the effects of energetic masking increase substantially when additional interfering talkers are added to a multitalker stimulus at a fixed overall SNR.

The results in Fig. 6 also provide some insight into how many simultaneous talkers are necessary for a multitalker speech stimulus to produce the same amount of masking as a speech-shaped noise signal at the same overall SNR value. Comparing the multitalker line to the continuous noise line, we see that the multitalker stimulus generally produced less energetic masking than the noise interferer when it contained seven or fewer simultaneous talkers, but that it actually produced slightly more energetic masking than the noise interferer when it contained nine or more simultaneous talkers. This result suggests that roughly six to eight interfering talkers are required to fill in the gaps in a multitalker stimulus to the point that it will produce the same amount of energetic masking as a continuous noise signal. Note that this is roughly consistent with the results of Simpson and Cooke (2005), which showed that the amount of energetic masking caused by a babble-modulated noise increased sharply as the number of competing talkers used to generate the babble increased from 1 to 6 but began to level off as the number of talkers increased beyond 6. However, Simpson and Cooke (2005) also showed that as many as 512 actual speech signals (as opposed to modulated noises) were required to produce a masker equivalent to speech-shaped noise, suggesting that the composition of the audible portion of the masker plays an important role in determining how well listeners can extract a target speech signal from a complex acoustic mixture. This result provides further evidence that performance in multitalker tasks is limited by the ability to extract usable target information from a mixture containing similar masker information than by a complete loss of target information due to spectro-temporal overlap with a more powerful masker.

V. CONCLUDING REMARKS

The experiments reported in this study extend our previous investigation where ITFS was introduced to remove or largely reduce informational masking effects (Brungart *et al.*, 2006). In Experiment 1, we have applied ITFS to multitalker mixtures where interfering talkers are either the same as or different from the target talker and the number of competing talkers is systemically varied. As in the previous study, the results show that performance is almost perfect when the ITFS method is applied to a multitalker signal with a TMR of 0 dB. As the effective TMR of the stimulus decreases below 0 dB, performance does eventually decrease. However, the vocal similarity of the target and masking signals had relatively little effect on performance for the ITFS stimuli. Changing from a stimulus with interfering phrases spoken by the target talker to a stimulus with masking phrases spoken by different same-sex talkers than the target talker resulted in a large release from non-energetic masking (as indicated by the large improvement in performance in Fig. 2 for ITFS-processed stimuli at LC values less than −6 dB) but only about a 1.19 dB release from energetic masking (as indicated by the relatively small improvement in performance in Fig. 2 for LC values greater than 0 dB). Changing from same-sex interfering talkers to different-sex interfering talkers produced only an additional 0.93 dB release from energetic masking. Thus it seems that energetic masking due to spectro-temporal overlap between the target and masking voices can account for only a tiny fraction of the release from masking that occurs when a same-sex masking voice is replaced with a voice spoken by a different-sex interfering talker.

Conceptually, the non-energetic masking values shown in Fig. 5 are closely related to the notion of informational

masking as it has been applied in the context of multitalker listening (Brungart, 2001; Carhart and Tillman, 1969; Freyman *et al.*, 1999; Kidd *et al.*, 1998; Pollack, 1975). In the context of ITFS, non-energetic masking refers to the release from masking that occurs when the acoustically detectable elements of the target signal are preserved but the acoustically detectable components of the masker are artificially removed from the stimulus. Thus, one would expect non-energetic masking to increase in cases where the target and masker are perceptually similar, and thus hard to distinguish from one another, and decrease in cases where the target and masker are perceptually different. In the most extreme case of masker dissimilarity, where the target signal is speech and the masker is Gaussian noise, Brungart *et al.* (2006) showed that ITFS processing resulted in only a 2–5 dB release from masking. In contrast, the application of ITFS processing in this experiment always produced at least a 12 dB release from masking. This suggests that non-energetic masking effects contribute substantially to the segregation of speech from a speech masker even when that masker is easily distinguished from the target speech by an obvious difference in talker sex. The results also showed a 9–14 dB release from non-energetic masking when switching from same-talker maskers to different-sex maskers. Again, these results are consistent with the notion that informational masking increases when target and masking voices are made progressively more similar to one another.

In Experiment 2, the ITFS technique was applied to a speech signal that was masked by up to 19 simultaneous competing speech signals. Remarkably, the results show that listeners could still perform well above chance when the ITFS procedure was applied to a mixture with 18 simultaneous interfering speech signals, even when all the competing phrases were spoken by the same talker. In contrast, performance was near chance (8%) with just four competing talkers when no ITFS processing was applied. These results show that there are enough target-dominated T-F regions to extract some information about a target speech signal even in extremely dense acoustic mixtures, and suggest that poor performance in multitalker tasks is probably more due to the inability to successfully identify and segregate the target acoustic information in the mixture than to a complete loss of all acoustic information related to the target speech.

ACKNOWLEDGMENTS

This research was supported in part by an AFRL grant via Veridian and an AFOSR grant (LRIR HE-01-COR-01). We thank Y. Li and Z. Jin for their assistance in figure preparation, and Richard Freyman for his helpful comments in the review process.

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Assmann, P. F., and Summerfield, A. Q. (**1990**). "Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.

Arbogast, T., Mason, C., and Kidd, G. (**2002**). "The effect of spatial separation on information and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Bird, J., and Darwin, C. J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London), pp. 263–269.

Bolia, R., Nelson, W. T., Ericson, M., and Simpson, B. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

Bronkhorst, A., and Plomp, R. (**1992**). "Effects of multiple speechlike maskers on binaural speech recognitions in normal and impaired listening," J. Acoust. Soc. Am. **92**, 3132–3139.

Brungart, D. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D. S., and Simpson, B. D. (**2002**). "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal," J. Acoust. Soc. Am. **112**, 664–676.

Brungart, D., Simpson, B. D., Ericson, M., and Scott, K. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Cahart, R., and Tillman, T. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.

Cavallini, F. (**1993**). "Fitting a logistic curve to data," Coll. Math. J. **24**, 247–253.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

Drullman, R. (**1995**). "Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796–1798.

Festen, J., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Freyman, R., Helfer, K., McCall, D., and Clifton, R. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3587.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.

Freyman, R. L., Helfer, K., and Balakrishnan, U. (**2007**). "Variability and uncertainty in masking by competing speech," J. Acoust. Soc. Am. **121**, 1040–1046.

Kidd, G., Mason, C., Rohtla, T., and Deliwala, P. (**1998**). "Release from informational masking due to the spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**, 422–431.

Li, N., and Loizou, P. C. (**2007**). "Factors influencing glimpsing of speech in noise," J. Acoust. Soc. Am. **122**, 1165–1172.

Li, Y., and Wang, D. L. (**2009**). "On the optimality of ideal binary time-frequency masks," Speech Commun. **51**, 230–239.

Miller, G. (**1947**). "Sensitivity to changes in the intensity of white Gaussian noise and its relation to masking and loudness," J. Acoust. Soc. Am. **191**, 609–619.

Moore, T. (**1981**). "Voice communication jamming research," in AGARD Conference Proceedings 331: Aural Communication in Aviation, Neuilly-SurSeine, France, pp. 2:1–2:6.

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). "SVOS final report, Part B: Implementing a gammatone filterbank," Report No. 2341, MRC Applied Psychology Unit, Cambridge, UK.

Peterson, G. H., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Pollack, I. (**1975**). "Auditory informational masking," J. Acoust. Soc. Am. **57**(S1), S5.

Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.

Simpson, S., and Cooke, M. (**2005**). "Consonant identification in N-talker babble is a nonmonotonic function of N," J. Acoust. Soc. Am. **118**, 2775–2778.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.

Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley, New York/IEEE, Hoboken, NJ.

Wightman, F. L., and Kistler, D. J. (**2005**). "Informational masking of speech in children: Effects of ipsilateral and contralateral distracters," J. Acoust. Soc. Am. **118**, 3164–3176.